IMPLEMENTASI METODE *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST* UNTUK DATASET TIDAK SEIMBANG (STUDI KASUS: KLASIFIKASI KEBANGKRUTAN PERUSAHAAN)

SKRIPSI

RESKIANTY

H071171509

PROGRAM STUDI SISTEM INFORMASI DEPARTEMEN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2022

IMPLEMENTASI METODE *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST* UNTUK DATASET TIDAK SEIMBANG (STUDI KASUS: KLASIFIKASI KEBANGKRUTAN PERUSAHAAN)

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains pada Program Studi Sistem Informasi Departemen Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin

RESKIANTY

H071171509

PROGRAM STUDI SISTEM INFORMASI DEPARTEMEN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2022

PERNYATAAN KEASLIAN

PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama	: Reskianty
NIM	: H071171509
Program Studi	: Sistem Informasi
Jenjang	: S1

Menyatakan dengan ini bahwa karya tulisan saya berjudul

IMPLEMENTASI METODE *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST* UNTUK DATASET TIDAK SEIMBANG (STUDI KASUS: KLASIFIKASI KEBANGKRUTAN PERUSAHAAN)

adalah karya tulisan saya sendiri dan bukan merupakan pengambilan alihan tulisan orang lain dan belum pernah dipublikasikan dalam bentuk apapun.

Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Makassar, 4 Juli 2022

Menyatakan,

METERAI TEMPEL

Reskianty

NIM: H071171509

iii

HALAMAN PERSETUJUAN PEMBIMBING

IMPLEMENTASI METODE *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST* UNTUK DATASET TIDAK SEIMBANG (STUDI KASUS: KLASIFIKASI KEBANGKRUTAN PERUSAHAAN)

Disusun dan diajukan oleh

RESKIANTY

H071171509

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Sistem Informasi Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin pada tanggal 4 Juli 2022 dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama	**Pembimbing Pertama**
Dr.Eng. Armin Lawi, S.Si., M.Eng.	**Andi Muh. Amil Siddik, S.Si., M.Si.**
NIP: 19720423 199512 1 001	NIP: 19911003 201903 1 015

Ketua Program Studi,

Dr. Muhammad Hasbi, M.Sc.

NIP: 19630720 198903 1 003

iv

HALAMAN PENGESAHAN

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh:

Nama	: Reskianty
NIM	: H071171509
Program Studi	: Sistem Informasi
Judul Skripsi	: Implementasi Metode *Support Vector Machine* dan *Random Forest* untuk Dataset Tidak Seimbang (Studi Kasus: Klasifikasi Kebangkrutan Perusahaan)

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Komputer pada Program Studi Sistem Informasi Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

DEWAN PENGUJI

		Tanda tangan
Ketua	: Dr.Eng. Armin Lawi, S.Si., M.Eng.	(.....)
Sekretaris	: Andi Muh. Amil Siddik, S.Si., M.Si.	(.....)
Anggota	: Andi Muhammad Anwar, S.Si., M.Si.	(.....)
Anggota	: Muhammad Sadno, S.Si., M.Si.	(.....)

Ditetapkan di : Makassar

Tanggal : 4 Juli 2022

v

KATA PENGANTAR

Segala puji dan syukur penulis panjatkan kepada Allah SWT Sang Maha Segalanya, tak lupa juga shalawat dan salam semoga senantiasa tercurahkan kepada junjungan kita Nabi Muhammad SAW atas seluruh curahat rahmat dan hidayah-Nya sehingga penulis mampu menyelesaikan skripsi yang berjudul "Implementasi Metode *Support Vector Machine* dan *Random Forest* untuk Dataset Tidak Seimbang (Studi Kasus: Klasifikasi Kebangkrutan Perusahaan)". Skripsi ini ditulis dalam rangka memenuhi syarat untuk mencapai gelar sarjana pada Program Studi Sistem Informasi Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin Makassar.

Dalam penyelesaian studi dan penulisan skripsi ini penulis menyadari bahwa banyak memperoleh ajaran dan kendala, namun berkat bantuan bimbingan, kerjasama dari berbagai pihak, dan berkah dari Allah SWT sehingga kendala-kendala yang dihadapi tersebut dapat diatasi. Untuk itu penulis menyampaikan penghargaan dan terima kasih yang tak terhingga kepada semua pihak yang telah membantu dalam proses penulisan skripsi ini, diantaranya adalah:

1. Rektor Universitas Hasanuddin, bapak **Prof. Dr. Ir. Jamaluddin Jompa, M.Sc.** beserta jajarannya.
2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam, **Dr.Eng. Amiruddin** beserta seluruh jajarannya.
3. Ketua Departemen Matematika FMIPA, **Prof. Dr. Nurdin, S.Si., M.Si.**, dan juga **Drs. Muhammad Hasbi, M.Sc.** sebagai ketua Program Studi Sistem Informasi Universitas Hasanuddin.
4. Bapak **Dr.Eng. Armin Lawi, S.Si., M.Eng.** sebagai pembimbing utama yang telah banyak memberikan arahan, ide, dan motivasi kepada penulis dalam banyak hal.
5. Bapak **Andi Muh. Amil Siddik, S.Si., M.Si.** sebagai pembimbing pertama yang telah memberikan masukan dan arahan kepada penulis.
6. Bapak **Andi Muhammad Anwar, S.Si., M.Si.** dan Bapak **Supri Bin Hj. Amir, S.Si., M.Eng.** sebagai tim penguji atas saran dan masukan pada penelitian yang telah dilakukan oleh penulis.

7. Bapak/Ibu Dosen Departemen Matematika Unhas yang telah memberikan ilmu kepada penulis selama menjadi mahasiswa di Departemen Matematika, dan seluruh staf Departemen Matematika dan Ilmu Komputer Unhas yang telah membantu penulis dalam urusan berkas administrasi.
8. Terima kasih kepada kedua orang tua ayahanda **Asri** dan ibunda **Armawati Sanusi**, untuk beliaulah skripsi ini dipersembahkan. Terima kasih yang telah membesarkan dengan penuh cinta dan kasih sayang, memberikan doa, motivasi, semangat, dukungan, dan berjuang hingga penulis mencapai perguruan tinggi. Kesuksesan dan segala hal baik yang kedepannya akan penulis dapatkan adalah karena untuk kalian berdua.
9. Teruntuk keluarga penulis lainnya, kakak **Nur Aisah**, adik **Fitri Amalia**, adik **Afdhal Gilang Aditya**, tante **Muliana**, om **Saparuddin**, sepupu **Mifthaul Jannah**, adik sepupu **Syawaliyah Fitri**, sepupu **Andriana**, **Mhuliana**, terima kasih selalu percaya pada mimpi-mimpi penulis, kalian adalah yang terbaik dan panutan sejak bayi. Kalian juga yang selalu mendukung, memberi semangat, dukungan moral, nasehat, doa, dan bisa memberikan kekuatan bagi penulis sehingga berada di titik ini.
10. Sahabat-sahabat **Popal**, **Ayu**, **Yustika**, **Nunu**, **Amel**, **Ardi**, **Farhan**, **Denny**, **Adhan**, **Alfandy**, **Irham**, **Dani**, yang telah menemani dari awal kuliah hingga saat ini, memberikan kebahagiaan di masa perkuliahan penulis, berbagi suka dan duka di kala bersama, menjadikan panutan dalam berteman, mampu mengajarkan apa arti dari kehidupan bersosial. Terima kasih banyak sahabat-sahabat **Popal** sudah menemani dan memberi bantuan penulis selama perkuliahan hingga saat ini.
11. Sahabat seperjuangan **Triana Rahayu**, **Atika**, **Husnaeni**, **Nurfadillah**, **Maylan**, **Nur Azizah**, **Nursyahputri Nasution**, para pendukung setia yang memberikan canda gurau, menghibur di saat sedih, menyalurkan hal-hal positif sifat yang membangun sehingga selalu menemani agar bisa semangat menyelesaikan skripsi ini.
12. Teman-teman **Program Studi Ilmu Komputer 2017** yang telah berjuang bersama dalam suka dan duka selama ini.

13. Teruntuk sobat gurun **Muthia Amanah Arum** dan **Ayu Naseer**, mereka adalah kacamata penulis, pendukung dari awal hingga saat ini, yang selalu setia dan sabar menghadapi penulis, memberikan arahan agar penulis semangat menyelesaikan skripsi ini.
14. Teruntuk orang kesayangan **Sahrul Ramadhan**, dia adalah motivator kedua setelah keluarga, terima kasih karena selalu ada di kala suka dan duka, penasehat terbaik, selalu bisa memberikan bantuan, dukungan penyemangat, ajakan liburan, arahan untuk masa depan, kebahagiaan yang terus menerus, dan selalu mendoakan agar penulis mampu menyelesaikan skripsi ini dengan baik.

Sebagai manusia biasa penulis menyadari penyusunan skripsi ini jauh dari kata sempurna karena keterbatasan kemampuan dan ilmu pengetahuan yang dimiliki oleh penulis. Oleh karenanya atas kesalahan dan kekurangan dalam penulisan skripsi ini, penulis memohon maaf dan bersedia menerima kritikan yang membangun.

Terakhir, penulis berharap semoga Tuhan Yang Maha Esa memberikan karunia rahmat dan hidayah-Nya kepada mereka semua. Semoga skripsi ini dapat bermanfaat bagi siapa saja yang membacanya.

Makassar, 4 Juli 2022

Reskianty

PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Hasanuddin, saya yang bertanda tangan di bawah ini:

Nama	: Reskianty
NIM	: H071171509
Program Studi	: Sistem Informasi
Departemen	: Matematika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam
Jenis karya	: Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Hasanuddin **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty-Free Right*)** atas karya ilmiah saya yang berjudul:

Implementasi Metode *Support Vector Machine* dan *Random Forest* untuk Dataset Tidak Seimbang (Studi Kasus: Klasifikasi Kebangkrutan Perusahaan)

beserta perangkat yang ada (jika diperlukan). Terkait dengan hal di atas, maka pihak universitas berhak menyimpan, mengalih-media/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di Makassar pada tanggal 2 Juli 2022

Yang menyatakan

(Reskianty)

ABSTRAK

Untuk mengantisipasi kemungkinan perusahaan akan mengalami kebangkrutan, perusahaan dapat melakukan prediksi kebangkrutan. Pada penelitian ini, akan digunakan dua metode yaitu *Support Vector Machine* dan *Random Forest*, namun data yang digunakan pada penelitian ini dikategorikan sebagai data tidak seimbang (*imbalanced data*) karena rasio data bangkrut dengan data tidak bangkrut tidak sama, sehingga perlu dilakukan *resampling* untuk menangani hal tersebut. Tahapan penelitian terbagi menjadi beberapa bagian, yakni tahap eksplorasi dan *preprocessing* data, pembagian data, *resampling* data, pembangunan model *classifier*, serta analisa terhadap hasil model *classifier*. Dari penelitian ini, diketahui bahwa filter SMOTE dalam data tidak seimbang memberikan pengaruh pada model *Machine Learning* yang diujikan, dalam hal ini *Support Vector Machine* dan *Random Forest*. Pengaruh yang dihasilkan juga bergantung pada pemilihan besar persen filter, karena kondisi data yang berbeda-beda tentunya memerlukan besaran persen filter yang berbeda pula. Dengan melihat hasil penelitian, dapat disimpulkan bahwa model *Random Forest* lebih baik dalam memprediksi kebangkrutan perusahaan dibandingkan dengan model *Support Vector Machine*.

Kata kunci: bangkrut, *Support Vector Machine, Random Forest, imbalanced data*.

ABSTRACT

To anticipate the possibility of the company going bankrupt, the company can predict bankruptcy. In this study, two methods will be used, namely Support Vector Machine and Random Forest, but the data used in this study is categorized as unbalanced data because the ratio of bankrupt data to non-bankrupt data is not the same, so resampling is necessary to deal with this problem. the. The research stages are divided into several parts, namely the exploration and data preprocessing stages, data sharing, data resampling, classifier model development, and analysis of the classifier model results. From this research, it is known that the SMOTE filter in the unbalanced data affects the Machine Learning model being tested, in this case, the Support Vector Machine and Random Forest. The effect also depends on the selection of the percent filter size, because different data conditions certainly require different filter percent sizes. By looking at the results of the study, it can be concluded that the Random Forest model is better at predicting corporate bankruptcy than the Support Vector Machine model.

Keywords: *bankrupt, Support Vector Machine, Random Forest, imbalanced data*.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1. Latar Belakang

Saat ini pandemi COVID-19 telah berdampak pada perubahan tatanan kehidupan sosial serta menurunnya kinerja ekonomi di sebagian besar negara di dunia, tak terkecuali di Indonesia; salah satu dampaknya yaitu kebangkrutan. Kebangkrutan berasal dari kata dasar bangkrut, yang menurut Kamus Besar Bahasa Indonesia, bangkrut adalah menderita kerugian besar hingga jatuh (tentang perusahaan, toko, dan sebagainya); gulung tikar. Kebangkrutan merupakan masalah yang harus diwaspadai oleh perusahaan, yang menurut Kamus Besar Bahasa Indonesia adalah perihal (keadaan) bangkrut dari perusahaan karena tidak mampu membayar utang-utangnya dan sebagainya. Dalam dunia perekonomian, kebangkrutan biasanya diartikan sebagai kegagalan perusahaan dalam menjalankan operasi perusahaan untuk menghasilkan laba (Suryani, 2011).

Untuk mengantisipasi kemungkinan perusahaan akan mengalami kebangkrutan, perusahaan dapat melakukan prediksi kebangkrutan. Prediksi kebangkrutan adalah tugas yang sangat penting bagi banyak lembaga keuangan terkait. Secara umum, tujuannya adalah untuk memprediksi kemungkinan bahwa suatu perusahaan akan bangkrut. Melalui prediksi tersebut dapat diketahui apakah perusahaan berada dalam kondisi operasional sehat atau mengarah pada kebangkrutan. Dalam menganalisa prediksi kebangkrutan, dibutuhkan model prediksi yang efektif untuk membuat keputusan pergerakan keuangan perusahaan yang tepat.

Metode-metode yang digunakan dalam memprediksi kebangkrutan juga beragam, mulai dari metode dari segi analisis statistik maupun dari segi analisis berbasis komputer. Analisis statistik contohnya seperti regresi linear, regresi logistik, analisis diskriminan, dan sebagainya, sedangkan analisis komputer contohnya seperti *Machine Learning* (ML), *Artificial Neural Network* (ANN), *trait recognition*, dan sebagainya.

Pada penelitian ini, akan digunakan dua metode yaitu *Support Vector Machine* dan *Random Forest* untuk memprediksi dan mengklasifikasi perusahaan yang mengalami kebangkrutan. Namun data yang digunakan pada penelitian ini

dikategorikan sebagai data tidak seimbang (*imbalanced data*) karena rasio data bangkrut dengan data tidak bangkrut tidak sama, sehingga perlu dilakukan *resampling* untuk menangani hal tersebut. Kemudian hasil dari kedua metode yang digunakan akan dibandingkan untuk mengetahui metode mana yang lebih baik performanya. Selain itu juga akan dibandingkan akurasi dari data sebelum dan setelah dilakukan *resampling*. Penelitian yang telah dijelaskan penulis sebelumnya akan ditulis pada skripsi berjudul "Implementasi Metode *Support Vector Machine* dan *Random Forest* untuk Dataset Tidak Seimbang (Studi Kasus: Klasifikasi Kebangkrutan Perusahaan)".

1.2. Rumusan Masalah

Berdasarkan latar belakang di atas, dapat dirumuskan beberapa masalah sebagai berikut:

1) Bagaimana melakukan eksplorasi dan *preprocessing* pada dataset tidak seimbang?
2) Bagaimana membangun model prediksi (*classifier*) kebangkrutan perusahaan dengan metode *Support Vector Machine* dan *Random Forest*?
3) Bagaimana pengaruh *resampling* dalam menangani dataset tidak seimbang?
4) Bagaimana analisis kinerja metode *Support Vector Machine* dan *Random Forest* dalam memprediksi kebangkrutan perusahaan?

1.3. Tujuan Penelitian

Dengan memperhatikan latar belakang dan rumusan masalah di atas, maka tujuan penelitian ini adalah:

1) Untuk mengetahui bagaimana melakukan eksplorasi dan *preprocessing* pada dataset tidak seimbang.
2) Untuk mengetahui bagaimana membangun model prediksi (*classifier*) kebangkrutan perusahaan menggunakan metode *Support Vector Machine* dan *Random Forest*.

3) Untuk mengetahui bagaimana pengaruh *resampling* dalam menangani dataset tidak seimbang.
4) Untuk mengetahui bagaimana analisis kinerja metode *Support Vector Machine* dan *Random Forest* dalam memprediksi kebangkrutan perusahaan.

1.4. Manfaat Penelitian

Manfaat penelitian dalam penulisan proposal ini adalah:

1) Dapat mengetahui proses eksplorasi dan *preprocessing* pada dataset tidak seimbang.
2) Dapat mengetahui efektivitas model prediksi kebangkrutan menggunakan metode *Support Vector Machine* dan *Random Forest*.
3) Dapat mengetahui pengaruh resampling pada dataset tidak seimbang.
4) Dapat digunakan sebagai model prediksi dasar untuk penelitian masa depan.

1.5. Batasan Masalah

Adapun batasan masalah yang diteliti untuk mencegah pembahasan yang terlalu luas, yaitu:

1) Merupakan masalah penelitian tentang penerapan *Artificial Intelligence*, khususnya cabang *Machine Learning*.
2) Metode *resampling* yang digunakan adalah metode SMOTE.
3) Pencapaian akhir penelitian merupakan hasil analisis dari model prediksi *Support Vector Machine* dan *Random Forest*.

BAB II
TINJAUAN PUSTAKA

2.1. Kebangkrutan

Kebangkrutan diartikan sebagai ketidakmampuan perusahaan untuk membayar kewajiban keuangannya pada saat jatuh tempo yang menyebabkan kebangkrutan atau kesulitan likuiditas yang mungkin sebagai awal kebangkrutan. Suatu perusahaan dianggap mengalami kebangkrutan atau kegagalan keuangan karena tingkat pengembalian yang diperoleh perusahaan lebih kecil dari total biaya yang harus dikeluarkannya dalam jangka panjang. Kesulitan keuangan yang terus menerus dihadapi perusahaan karena biaya yang dikeluarkan lebih besar dari pendapatannya akan mengancam kelangsungan usaha perusahaan dalam jangka panjang. Akumulasi kesulitan mengelola keuangan dalam jangka panjang akan mengakibatkan nilai aset yang lebih kecil dibandingkan dengan kewajiban totalnya (Rudianto, 2013).

Faktor-faktor kebangkrutan terdiri dari dua macam, yaitu kondisi internal perusahaan, dimana terlalu besarnya kredit yang diberikan kepada debitur/langganan, manajemen yang tidak efisien, kesalahan dalam menetapkan harga jual, pengelolaan utang piutang yang kurang memadai, struktur biaya (produksi, administrasi, pemasaran, dan finansial) yang tinggi, dan faktor-faktor lain. Selain itu, ada pula kondisi eksternal perusahaan yang bersifat umum, misalnya faktor politik, ekonomi, sosial dan budaya, penggunaan teknologi yang keliru, dan faktor-faktor lain (Munawir, 2010).

Ada beberapa hal yang dapat dijadikan alasan untuk menyatakan bahwa perusahaan tersebut berada dalam kondisi kesulitan keuangan, antara lain terjadinya penurunan aset, penurunan penjualan, perolehan laba dan profitabilitas yang semakin rendah, berkurangnya modal kerja, dan tingkat hutang yang semakin tinggi (Hani, 2015).

2.2. Data Mining

Data Mining merupakan tahapan untuk menemukan korelasi, pola, dan tren terhadap suatu hal berdasarkan sumber data yang ada (Liu, 2007). Dengan kata lain, *data mining* adalah proses untuk mencari sebuah informasi atau pengetahuan yang

belum diketahui dari data yang ada, kemudian menghasilkan informasi yang berguna untuk keperluan pengolahan data selanjutnya. Rangkaian proses data *mining* secara keseluruhan meliputi seleksi data, *preprocessing*, transformasi data, pencarian informasi, dan evaluasi atau interpretasinya (Kurniawansyah, 2018).

Seleksi data adalah tahap awal proses *mining* informasi, data dipilih berdasarkan masalah yang ingin dipecahkan. *Preprocessing* meliputi *data cleaning* atau proses menangani data berbentuk missing value, feature scaling untuk menormalkan nilai *variable independent* pada dataset, dan *feature selection* yang berguna mereduksi dimensi dataset tanpa harus menghilangkan keseluruhan bobot informasi pada dataset. Tahap transformasi berguna untuk mengubah data ke dalam sebuah format yang dapat digunakan secara *cross-application* atau independen terhadap media pengolahan yang digunakan. Kemudian ada tahap evaluasi dan interpretasi yang bertujuan memberikan pemahaman atau *insight* secara jelas.

2.2.1. Feature Scaling

Feature scaling merupakan metode untuk menormalkan *feature* atau atribut independen pada dataset, yang bertujuan membuat skala yang baku untuk setiap *feature* (Larose, 2014). Terdapat dua cara *scaling* yang biasa digunakan, standarisasi dan normalisasi.

$$x_{normalize} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2.1}$$

$$x_{normalize} = \frac{x - \text{mean}(x)}{\text{standard dev}(x)} \tag{2.2}$$

Normalisasi atau yang dikenal sebagai *min-max scaling* merupakan teknik *scaling* data dengan cara mentransformasikan skala data yang ada pada batas 0 sampai 1. Sedangkan *scaling* dengan metode standarisasi mentransformasikan skala data untuk terpusat pada standar deviasi atribut.

2.2.2. SMOTE

SMOTE atau *Synthetic Minority Oversampling Technique* adalah teknik *oversampling* terpopuler yang diproposalkan oleh Chawla (Chawla,

dkk., 2002) pada tahun 2002. Teknik ini membuat data tiruan atau sintetik berdasarkan tetangga-tetangga terdekat dari sampel kelas minoritas.

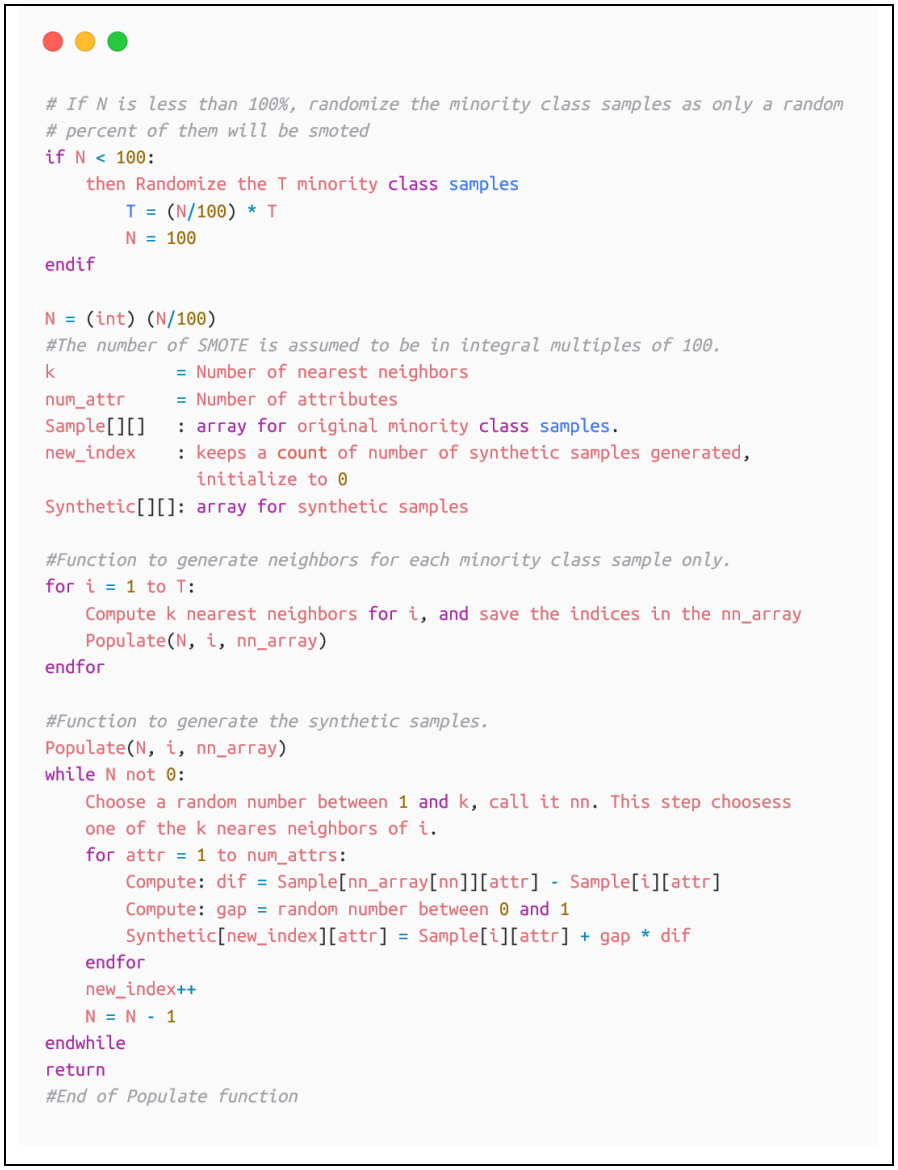
Metode SMOTE bekerja dengan membuat replikasi dari kelas minoritas. Kelas minoritas di *oversampling* dengan cara mengambil setiap sampel kelas minoritas dan memperkenalkan contoh sintetis di sepanjang garis segmen yang bergabung dengan setiap/semua *k-nearest neighbors* kelas minoritas. Setelah itu akan dibuat data sintetis sebanyak persentase duplikasi yang diinginkan antara data minor dengan *k-nearest neighbors* yang dipilih secara acak.

Teknik ini dimulai dengan menentukan *n* data yang akan dibuat untuk setiap data pada kelas minoritas dalam dataset *D*. Kemudian untuk setiap data kelas minoritas $M_{i,c}$, pilih n tetangga secara acak dari *k* tetangga terdekat data tersebut, di mana *i = {1,2,...m}* dengan *m* adalah jumlah data pada kelas minoritas dan $c \in C$, dimana $C = \{c_1, c_2, \dots, c_z\}$ adalah fitur-fitur pada *D*. Lalu untuk setiap fitur *c* pada $M_{i,c}$, hitung jarak Euclid *d* antara $M_{i,c}$ dengan salah satu tetangga $T_{i,c}$, di mana s adalah bilangan acak *s* = {1, 2, …, *k*} dari *k* tetangga terdekat $M_{i,c}$. Kemudian suatu bilangan acak *g* = [0, 1] ditentukan. Data tiruan dibuat berdasarkan:

$$S_{j,c} = M_{i,c} + (g + d) \tag{2.3}$$

di mana *j* ={1,2,.., *n*m*} bersifat *incremental* dan *S* adalah data kelas minoritas tiruan. Teknik ini membuat *n*m* data sintetik pada suatu titik dari jarak antara setiap fitur dari *M* dengan *T* (Chawla, dkk., 2002).

Algoritma SMOTE dapat dilihat pada gambar 2.1.



```
# If N is less than 100%, randomize the minority class samples as only a random
# percent of them will be smoted
if N < 100:
    then Randomize the T minority class samples
        T = (N/100) * T
        N = 100
endif

N = (int) (N/100)
#The number of SMOTE is assumed to be in integral multiples of 100.
k            = Number of nearest neighbors
num_attr     = Number of attributes
Sample[][]   : array for original minority class samples.
new_index    : keeps a count of number of synthetic samples generated,
               initialize to 0
Synthetic[][]: array for synthetic samples

#Function to generate neighbors for each minority class sample only.
for i = 1 to T:
    Compute k nearest neighbors for i, and save the indices in the nn_array
    Populate(N, i, nn_array)
endfor

#Function to generate the synthetic samples.
Populate(N, i, nn_array)
while N not 0:
    Choose a random number between 1 and k, call it nn. This step choosess
    one of the k neares neighbors of i.
    for attr = 1 to num_attrs:
        Compute: dif = Sample[nn_array[nn]][attr] - Sample[i][attr]
        Compute: gap = random number between 0 and 1
        Synthetic[new_index][attr] = Sample[i][attr] + gap * dif
    endfor
    new_index++
    N = N - 1
endwhile
return
#End of Populate function
```

Gambar 2.1. Pseudocode Algoritma SMOTE

Dari gambar 2.1 diatas, algoritma *SMOTE (T, N, k)* terdiri dari 3 parameter input yaitu:

T: jumlah sampel kelas minoritas

N: jumlah SMOTE dalam N%

k: jumlah *nearest neighbors*

Yang menghasilkan output berupa (*N*/100) * *T* (Chawla, dkk., 2002).

2.3. *Machine Learning*

Machine Learning (ML) adalah proses *artificial* dimana suatu algoritma dikembangkan untuk melakukan iterasi pembelajaran secara repetitif dengan data

yang sudah ada, tanpa deprogram atau diperintahkan secara *hard-coded* (El Naqa & Murphy, 2015). Terdapat tiga jenis pembelajaran pada ML; *supervised learning*, *semi-supervised learning*, dan *unsupervised learning*.

Pembelajaran yang terawasi atau *supervised learning*, sebuah *classifier* membutuhkan suatu input tertentu dan menghubungkannya dengan suatu *output*. Dalam hal ini, kasus di mana tujuannya adalah mengklasifkasikan input data ke suatu kategori diskrit tertentu disebut klasifikasi, dan kasus di mana *output*-nya adalah suatu variabel kontinu disebut regresi (Alpaydin, 2020). Dalam pembelajaran tanpa pengawasan atau *unsupervised learning*, *classifier* diberi input berupa data latih, kemudian secara repetitif menemukan pola dari data tersebut. Kasus *unsupervised learning* yang tujuannya adalah mengelompokkan observasi-observasi yang mirip disebut *clustering*, dan penentuan distribusi pada data disebut estimasi kepadatan (Celebi & Aydin, 2016). Dalam *semi-supervised learning*, *classifier* menerima *input* secara terus menerus diikuti dengan pemilihan *decision* berdasarkan kondisi model dan dataset.

2.4. Klasifikasi

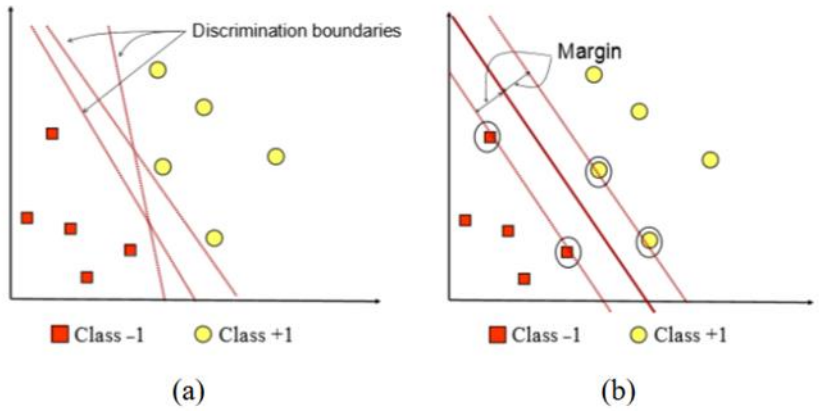
Klasifikasi merupakan bagian dari metode prediksi, namun yang diprediksi adalah label atau target dari sebuah *tuple* data. Dengan kata lain, tujuan klasifikasi adalah menerima suatu input sejenis dengan dataset awal, kemudian menentukan kelas target atau output-nya (Kotsiantis dkk., 2007).

Menurut Widiarti (2020), klasifikasi didasarkan oleh empat komponen, yaitu:

1. *Training dataset*, yakni data yang digunakan untuk melatih model dalam mengenali kelas-kelas berdasarkan variabel independen dan variabel dependen yang ada.
2. *Testing dataset*, yaitu data yang digunakan untuk menguji model *classifier* yang telah dilatih pada proses sebelumnya.
3. *Predictor*, adalah variabel independen, atau dapat disebut sebagai faktor yang mempengaruhi *target class* dari dataset.
4. *Target class*, merupakan variabel dependen yang dipengaruhi oleh *predictor* pada suatu dataset.

2.4.1. *Support Vector Machine*

Support Vector Machine (SVM) merupakan metode untuk menyelesaikan masalah statistika klasik, khususnya pada *task* klasifikasi dan prediksi. SVM sendiri dikembangkan menggunakan prinsip *linear classifier*. Namun seiring berkembangnya metodologi klasifikasi, sering pula dijumpai data yang bersifat non-linear, sehingga dikembangkan SVM dengan konsep *kernel* untuk pemetaan yang akurat (Hsu et al., 2003).



Gambar 2.2. Ilustrasi model SVM (Fachruddin, 2015)

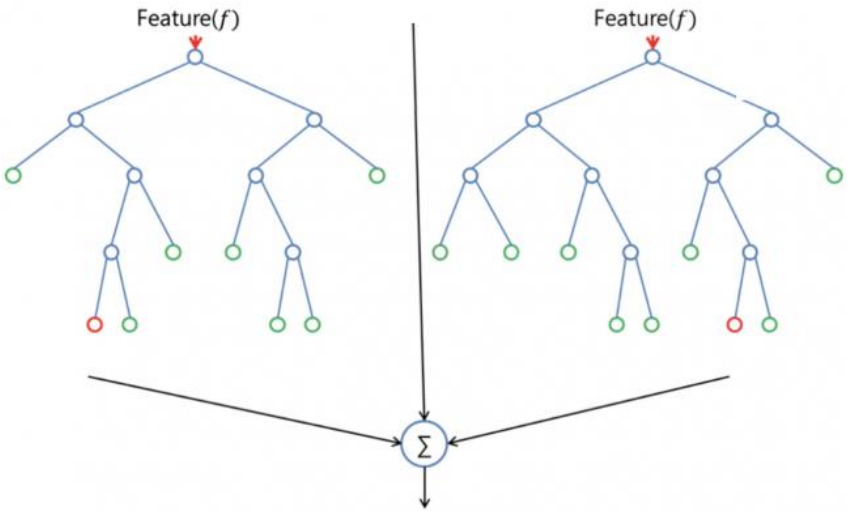
Konsep SVM adalah menemukan *hyperplane* pada ruang input, dan *hyperplane*-nya sendiri digunakan sebagai pemisah antara dua kelas pada ruang input. Pada gambar 2.2(a) diatas, tiga garis merah disebut sebagai *discriminant boundaries* atau garis pemisah yang dihasilkan untuk kelas yang ada, dan pattern yang paling dekat disebut sebagai *support vector*. Sedangkan garis tebal pada gambar 2.2(b) menunjukkan *hyperplane* terbaik antar kelas, karna terletak tepat diantara kedua kelas.

2.4.2. *Random Forest*

Random Forest (RF) adalah kombinasi *predictor* berbentuk *tree*, dan tiap *tree*-nya bergantung pada nilai suatu vektor acak berdasarkan sampling secara independen dengan distribusi yang setara (Breiman, 2001). Proses *randomization* untuk membentuk *tree* tidak hanya dilakukan pada data sampel saja, melainkan juga pada pengambilan variabel *predictor*-nya, hal ini akan menghasilkan kumpulan *classification tree* dengan ukuran dan bentuk yang berbeda-beda. Hasil yang diharapkan adalah suatu kumpulan

classification tree yang memiliki korelasi kecil antar *tree*-nya. Korelasi yang kecil akan menurunkan hasil kesalahan prediksi RF (Breiman, 2001).

Salah satu dari kelebihan RF adalah dapat digunakan baik untuk menyelesaikan masalah klasifikasi maupun masalah regresi, yang merupakan sistem *machine learning* saat ini. Selain itu, RF dikenal dapat digunakan pada *data imbalance* di *class* yang berbeda, khususnya untuk dataset yang berukuran besar (Paul dkk., 2018).



Gambar 2.3. Ilustrasi model Random Forest

Dari gambar 2.3, dapat dilihat bahwa RF adalah *classifier* yang terdiri dari beberapa *decision tree*, yang masing-masing berperan pada class tertentu. Dengan menggabungkan beberapa *tree* di RF dapat menghasilkan prediksi yang lebih baik (Devetyarov, 2010).

2.5. *Company Bankruptcy Prediction Data*

Company Bankruptcy Prediction Data (Liang et. al, 2021) adalah dataset yang menyajikan data prediksi kebangkrutan suatu perusahaan yang dikumpulkan melalui jurnal ekonomi Taiwan dari tahun 1999 sampai dengan tahun 2009, dan topik kebangkrutan perusahaan dinyatakan dalam regulasi bisnis Taiwan. CBPD mempunyai satu target dan 95 fitur atau atribut independen. Kemudian jumlah data pada CBPD ini berisikan 6819 baris data/*tuple*. Dataset ini akan dibagi menjadi dua bagian dengan rasio 7:3 atau untuk 70% data *training* dan 30% data *testing*, secara berurutan.

2.6. Ukuran Kinerja

Dalam pengukuran performa model klasifikasi, digunakan konsep *confusion matrix* yang menghasilkan skor dari prediksi model terhadap data testing. *Confusion matrix* merupakan tabel matriks dengan empat kombinasi kategori yang berbeda, berisikan nilai aktual dan nilai prediksi. Kombinasi yang ada antara lain a) *True Positive*, b) *True Negative*, c) *False Positive*, d) *False Negative*.

		Actual Class	
		Positive (**P**)	Negative (**N**)
Predicted Class	Positive (**P**)	True Positive (**TP**)	False Positive (**FP**)
	Negative (**N**)	False Negative (**FN**)	True Negative (**TN**)

Gambar 2.4. Ilustrasi Confusion Matrix (sumber: ml-science.com/confussion-matrix)

Dengan menggunakan tabel tersebut, dapat diukur performa dari model seperti *accuracy*, *precision*, dan nilai *recall* dengan persamaan sebagai berikut.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.4}$$

$$precision = \frac{TP}{TP + FP} \tag{2.5}$$

$$recall = \frac{TP}{TP + FN} \tag{2.6}$$

Accuracy merepresentasikan seberapa akurat model mengklasifikasikan tiap-tiap *tuple* data dengan benar. *Precision* menilai keakuratan data yang diminta dengan hasil prediksi model. Sedangkan *Recall* menggambarkan sensitivitas keberhasilan model dalam menemukan kembali informasi skor yang mungkin tercapai jika model memprediksi data dengan sempurna.